HATECAT-TR: A Hate Speech Span Detection and Categorization Dataset for Turkish

Hasan Kerem Şeker [1] and **Gökçe Uludoğan** [1] and **Pelin Önal**[2] and **Arzucan Özgür**[1]

[1]Department of Computer Engineering, Bogazici University, Istanbul, Turkey 34342
[2] Hrant Dink Foundation, Istanbul, Turkey 34373

hasan.seker@std.bogazici.edu.tr, gokce.uludogan@bogazici.edu.tr,
pelinonal@hrantdink.org, arzucan.ozgur@bogazici.edu.tr

Abstract

Hate speech on social media in Turkey remains a critical issue, frequently targeting minority groups. Effective moderation requires not only detecting hateful posts but also identifying the specific hateful expressions within them. To address this, we introduce HATECAT-TR, a span-annotated dataset of Turkish tweets, containing 4465 hateful spans across 2981 posts, each directed at one of eight minority groups. Annotations were created using a semi-automated approach, combining GPT-4o-generated spans with human expert review to ensure accuracy. Each hateful span is categorized into one of five discourse types, enabling a fine-grained analysis of the nature and intent behind hateful content. We frame span detection as binary and multi-class token classification tasks and utilize the state-of-the-art language models to establish a baseline performance for the new dataset. Our findings highlight the challenges of detecting and categorizing implicit hate speech, particularly when spans are subtle and highly contextual. The source code is available at github.com/boun-tabi/hatecat-tr and HATECAT-TR can be shared by complying with the terms of X upon contacting the authors.

⚠ **Warning:** This paper contains hate speech directed towards specific groups.

1 Introduction

Identifying hateful content is crucial for effective content moderation. While many automated systems detect hate speech, they often classify entire posts as hateful, lacking the granularity needed for informed decision-making. A recent study found that structured explanations improve moderator efficiency, reducing decision-making time by 7.4% (Calabrese et al., 2024). While explanations can take many forms, identifying and categorizing hateful spans provides an effective approach, helping moderators quickly assess which parts of a post are problematic and why. Consider the following example where the hateful spans are highlighted: "@username Bak Yunan tohumu, Feodalite denen bir şey vardı. Korkudan surların arkasına saklanmıştınız." (English translation: "@username Look, Greek seed. There was something called feudalism. You hid behind the walls out of fear."). Here, "Yunan tohumu" (Greek seed) falls under symbolization, as it reduces a group to a derogatory term, reinforcing negative stereotypes. Meanwhile, "Korkudan surların arkasına saklanmıştınız" (You hid behind the walls out of fear) is categorized as threat of enmity/war, as it implies historical cowardice and weakness, potentially inciting hostility. Instead of flagging the entire post, detecting and categorizing these spans allows moderators to distinguish the nature of hate speech, prioritize severe cases, ensure consistency in content moderation, and improve overall transparency in enforcing platform guidelines.

Beyond post-level classification, span detection has been explored to provide a more granular understanding of hate speech across various languages and domains. Studies have investigated span detection in English (Zhu et al., 2021; Mathew et al., 2021; Pavlopoulos et al., 2022; Masud et al., 2022; Zhou et al., 2023; Jafari et al., 2024), Korean (Jeong et al., 2022), and Vietnamese (Hoang et al., 2023), with applications in news, social media platforms, and extremist forums. These efforts demonstrate the importance of identifying specific hateful expressions to improve interpretability and moderation strategies.

Hate speech detection in Turkish has gained increasing attention, with various datasets and models developed to tackle hateful content (Çöltekin, 2020; Beyhan et al., 2022; Toraman et al., 2022). Shared tasks such as SIU2023 and HSD-2Lang have contributed benchmark datasets, advancing the field with transformer-based models and en-

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 25568–25579
November 4-9, 2025

semble approaches (Uludoğan et al., 2024; Arın et al., 2023). To our knowledge, the only study that addresses span-level detection in Turkish is JL-Hate (Büyükdemirci et al., 2024), which provides span-level annotations but lacks categorization and suffers from a limited dataset size (Büyükdemirci et al., 2024). These gaps underscore the need for a comprehensive Turkish dataset with detailed span annotations and categorical labels, enabling more explainable hate speech detection, and improved content moderation.

To address these gaps, we introduce HATECAT-TR, a span-annotated dataset of Turkish tweets, each targeting specific ethnic, religious, and other minority groups. Our dataset is the first Turkish dataset for hate speech span detection and categorization, offering a larger, more structured resource for hate speech research in Turkish. We construct the dataset using a semi-automated annotation process, combining GPT-4o generated spans with human expert review. Hateful spans are labeled with five distinct categories.

We approach span detection as both a binary token classification problem (identifying hateful spans) and a multi-class token classification problem (distinguishing between hate speech types). We establish baseline results using state-of-the-art language models. The comprehensive dataset with fine-grained annotations for hate speech spans and their categories as well as the developed baseline models will foster further research in this challenging but important area.

This research can be useful in several areas. For instance, in sociolinguistics, the detailed data on hate speech can help researchers understand how language and culture impact the way hate is expressed online. The span-level annotations also make it possible to build more transparent content moderation tools that not only detect hate speech but also explain why something is flagged. This can lead to more effective moderation and even help guide users toward avoiding hateful language in the future.

2 HATECAT-TR

Initial Dataset Identifying hateful content spans is essential for combating hate speech. While tweet-level detection determines if a tweet contains hate speech, span detection pinpoints the specific segments responsible, providing more granular insights. To support this task, we curated a dataset of tweets annotated with hate speech spans.

The dataset was sourced from a nongovernmental organization that has been working on tracking and analysis of hateful discourse in Turkish media. It consists of tweets labeled for hate speech presence and the specific target groups affected including the Armenian, Greek, Jewish, Arab, Immigrant/Refugee, LGBTI+, Alevi, and Kurdish groups. Part of this dataset has been previously used in the HSD2Lang Hate Speech Detection Challenge (Uludoğan et al., 2024). In this study, we provide an extended version of this data set annotated with hate speech spans and their categories.

Automated span annotation To automate span annotation, we employed the GPT-4o model (gpt-4o-2024-08-06) (Achiam et al., 2023), extracting hate speech spans within the tweets. We used a structured system message, provided in Appendix B, to guide GPT-4o in identifying and labeling hate speech spans. Each tweet was pre-assigned a single target during the prompting stage and processed individually using OpenAI's official API. The model was queried with the following input format: "Tweet id: [TWEET_ID], Target: [TARGET], Text: [CONTENT]". It was instructed to extract specific spans and classify them under one of the following five predefined categories.

- **Exclusive/Discriminatory Discourse**: Language that excludes or differentiates negatively based on group membership.
- **Exaggeration/Generalization/Attribution /Distortion**: Making unfair generalizations or attributing negative traits to a group.
- **Threat of Enmity/War/Attack/Murder /Harm**: Statements inciting violence or harm towards a group.
- **Symbolization**: Language that reinforces negative stereotypes through symbols or indirect references.
- **Swearing/Insult/Defamation/Dehumanization**: Direct insults or dehumanizing language targeting a specific group.

The five-category taxonomy used in this study was originally developed by the Hrant Dink Foundation through expert consultation and qualitative analysis of hate speech in Turkish media. It was later refined in collaboration with a multidisciplinary academic team and is actively used by the Foundation in its monitoring work (Foundation, 2019, 2025).

Our categorization aligns with prior work that

differentiates hate speech based on the nature and form of discourse. A comparable five-level scheme was introduced by Beyhan et al. (2022) and later adopted in the SIU2023-NST Hate Speech Detection Contest (Arın et al., 2023), with only minor differences in category definitions. While our taxonomy closely mirrors the SIU2023 scheme, it includes an additional category that explicitly captures exclusionary or discriminatory discourse. Similar discourse-based classification is also used by Sanguinetti et al. (2018), which proposed a four-level taxonomy for Italian-language content to capture escalating harmful discourse.

The automated span annotation approach was prone to errors, including minor modifications to the original tweet text and hallucinated spans that introduced unrelated content. To mitigate these issues, hallucinated spans were discarded, while minor textual differences were resolved using the edit distance metric. If a GPT-4o-annotated span differed by fewer than four characters from an existing span in the original tweet, the closest matching span in the tweet was retained.

Manual Review Additional refinement was required to ensure high-quality span annotations. We conducted a two-stage manual review of the generated spans, the details of which are provided in Appendix C and D. Two annotators manually reviewed the GPT-4o-generated spans, deleting, creating, or modifying spans to capture hate speech and assign appropriate categories accurately. Annotations with substantial disagreements regarding the hatefulness of spans or an Intersection over Union (IoU) score below 0.5 between the two annotators were re-evaluated by a third annotator.

Statistics After the re-evaluation process, the final dataset consists of 2981 tweets and 4465 hateful spans. Of the 2981 tweets annotated by GPT-4o, 1530 were accepted without modification, while 1451 were revised by human annotators. These revisions involved adjusting span boundaries, correcting category labels, or removing hallucinated content. Tables 1 and 4 present the dataset statistics for categories and target groups, respectively. Although the annotation framework supported multi-label spans, this was rarely observed; most spans were annotated with a single discourse type. On average hateful spans are 30.30 characters long with a 21.69 standard deviation. For all categories, there are tweets containing multiple instances of hateful spans within the same tweet. The average number of spans per tweet is 1.50.

Category	#Tweets	#Spans
Exclusive/Discriminatory Discourse	683	856
Exaggeration; Generalization; Attribution; Distortion	868	1104
Threat of Enmity; War; Attack; Murder; or Harm	1015	1396
Symbolization	700	800
Swearing; Insult; Defamation; Dehumanization	1015	1373

Table 1: Statistics based on hate speech categories.

Inter-Annotator Agreement Following prior work (Hoang et al., 2023), we assessed inter-annotator agreement using character-level Cohen's Kappa (McHugh, 2012) and character-level micro-averaged F1-score. Due to the class imbalance between negative and positive labels, Cohen's Kappa tends to be skewed and may not reliably reflect agreement on meaningful annotations. Therefore, we computed the F1-score by excluding negative labels, following Deleger et al. (2012), which suggest that this approach is more appropriate for evaluating inter-annotator agreement in NLP tagging tasks. In contrast, Cohen's Kappa is calculated over all character-level labels, including negatives. The results are summarized in Table 5 (Appendix F).

Annotators achieved a moderate Kappa score of 0.57 and a higher overall F1-score of 0.77, indicating strong agreement in identifying non-hateful spans but challenges in categorizing hateful spans (F1 = 0.50). Agreement was highest for the "Threat of Enmity; War; Attack; Murder; or Harm" category (F1 = 0.66), as it involves explicit harmful actions. In contrast, "Symbolization" was the most challenging (F1 = 0.29) due to its reliance on subtle negative stereotypes.

Cohen's Kappa penalizes disagreements in category labels even when annotators select the same span, which may underestimate agreement in subjective or ambiguous cases. To account for near matches in span annotations, we computed a relaxed version of Cohen's Kappa using a span-level evaluation protocol. In this approach, two annotated spans are considered a match if their Intersection over Union (IoU) exceeds a specified threshold. This relaxed metric tolerates minor discrepancies in span boundaries, focusing on whether annotators identified roughly the same region. As a result, the agreement score increases from 0.57 (standard Kappa) to 0.65 at 90% IoU, and further to 0.76 at 50% IoU, indicating that annotators often agree on the approximate region even if their exact span selections differ.

3 Experiments

We frame span detection and categorization as binary and multi-class token classification tasks, where each token is assigned a label indicating whether it is part of a hateful span and category. Due to the complexity of detecting hate speech and distinguishing between overlapping categories, we model categorization as a multiclass task by assigning the most salient category to each span. This reflects the dominant pattern of single-label spans in the data and simplifies the learning setup.

In the hate speech span detection task, BIO tagging is applied for encoder-based models (e.g., BERT variants). The **O** label marks tokens outside of hateful spans, **B-HATE** denotes the beginning of a hateful span, and **I-HATE** is used for tokens continuing the span. For the hate speech categorization task, we adopt an IO tagging scheme to simplify the process for encoder-based models. The label **O** is assigned to tokens outside any hateful span, while tokens within a span receive an **I-** prefix corresponding to the specific hate speech category (e.g., **I-Threat**).

For decoder-only LLMs, we do not follow a token classification framework and therefore do not use BIO or IO tagging. Instead, we used the same natural language prompt format as in our GPT-4o setup, shown in Listing 1 in Appendix B.

4 Results

We establish baseline results using two state-of-the-art Turkish language models, BERTurk (Schweter, 2020) with finetuning and TurkishBERTweet (Najafi and Varol, 2024), as well as multilingual models from the LLaMA (Grattafiori et al., 2024) and Gemma families (Team et al., 2024). Models were evaluated using span-level F1-scores. As shown in Table 2, BERTurk achieves the highest F1-scores on both tasks, followed by TurkishBERTweet. Both models outperform multilingual variants of LLaMA and Gemma evaluated under different strategies, including zero-shot, few-shot, and fine-tuning settings. However, all models score lower on the categorization task, indicating that this task is more difficult.[1]

Table 6 in Appendix provides a breakdown of F1-scores across hate speech categories. Both models perform best in the "Threat of Enmity; War; Attack; Murder; or Harm" category, with BERTurk reaching 0.44 and TurkishBERTweet achieving 0.35. This suggests that explicit threats and violent language are easier to classify compared to more nuanced hate speech types. In contrast, both models struggle with "Exclusive/Discriminatory Discourse" and "Exaggeration; Generalization; Attribution; Distortion". These results indicate that more implicit and context-dependent hate speech categories are difficult to detect and categorize. Similar trends are observed across multilingual models.

Task	Model	Precision	Recall	F1
Detection	**TurkishBERTweet**	0.50 ± 0.01	0.55 ± 0.02	0.53 ± 0.02
	BERTurk	0.57 ± 0.02	0.62 ± 0.01	0.59 ± 0.01
	Llama-3.3-70B Few Shot	0.44 ± 0.01	0.67 ± 0.01	0.53 ± 0.01
	Llama-3.3-70B Zero Shot	0.36 ± 0.01	0.71 ± 0.01	0.48 ± 0.01
	Llama-3.1-8B Finetuning	0.50 ± 0.04	0.21 ± 0.04	0.30 ± 0.03
	Gemma2-9B-Finetuning	0.62 ± 0.05	0.36 ± 0.02	0.46 ± 0.01
Categorization	**TurkishBERTweet**	0.24 ± 0.03	0.27 ± 0.02	0.25 ± 0.02
	BERTurk	0.29 ± 0.02	0.34 ± 0.02	0.31 ± 0.02
	Llama-3.3-70B Few Shot	0.18 ± 0.01	0.27 ± 0.01	0.22 ± 0.01
	Llama-3.3-70B Zero Shot	0.15 ± 0.01	0.27 ± 0.01	0.19 ± 0.05
	Llama-3.1-8B Finetuning	0.15 ± 0.02	0.07 ± 0.01	0.10 ± 0.02

Table 2: Performance comparison of the models on the hateful span detection and categorization tasks.

The confusion matrix in Figure 1 highlights key misclassification patterns: (1) Poor recognition of exclusive/discriminatory discourse, particularly in ethnic profiling statements such as "Bütün Türkiyeli siyasilere DNA ırk testi yapılarak açıklanmalı, kim Yahudi geni taşıyor, kim Yunan geni taşıyor, kim Ermeni geni taşıyor." ("A DNA ancestry test should be conducted on all politicians in Turkey, and the results should be publicly announced—who carries Jewish, Greek, or Armenian genes."). (2) Failures in detecting exaggeration and generalization, particularly in demographic-based fear-mongering, such as "Aynen Arapça öğren de, 5 sene içinde ülkenin yarısı Arap olunca Araplara da anlatırsın vatana sahip çıkmayı." ("Learn Arabic, so in five years, when half the country is Arab, you can explain to them how to protect the homeland."). (3) False negatives in threat-related speech, where phrases like "Yunan'ı denize döktük" ("We threw the Greeks into the sea") were overlooked due to historical or patriotic framing. (4) Missed swearing and dehumanization, particularly obfuscated insults ("pzvnk" for "pezevenk", "pimp") and indirect comparisons such as "Bunların bir Rum, bir Ermeni kadar bile insanlığı yok" ("They don't have even as much humanity as a Greek or an Armenian"). These findings suggest that the model requires improved handling

[1] We couldn't test Gemma on categorization because of the model hallucination creating invalid outputs as discussed in section G of the Appendix.

of implicit biases, coded language, and contextual nuance for more accurate classification.

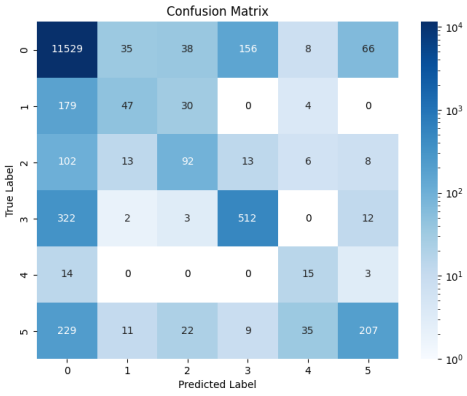

Figure 1: Hate speech categorization confusion matrix 0: Not hateful token 1: Exclusive/Discriminatory Discourse token 2: Exaggeration; Generalization; Attribution; Distortion token 3: Threat of Enmity; War; Attack; Murder; or Harm token 4: Symbolization token 5: Swearing; Insult; Defamation; Dehumanization token

5 Conclusion

We introduce HATECAT-TR, a Turkish tweet dataset annotated with hate speech spans and their categories. The dataset was created using automated GPT-4o annotations, followed by manual expert review, ensuring high-quality labeled spans across five hate speech categories.

Our evaluation demonstrates BERTurk's better performance over TurkishBERTweet for the span detection and categorization tasks. Both models perform best in detecting explicit threats, particularly in "Threat of Enmity; War; Attack; Murder; or Harm", but struggle with more implicit categories like "Exclusive/Discriminatory Discourse" and "Exaggeration; Generalization; Attribution; Distortion." Multilingual models, including variants of LLaMA and Gemma, perform noticeably worse than the Turkish models across all categories, highlighting the importance of language-specific pretraining for nuanced hate speech detection. Notably, the categorization task is challenging not only for the models but also for human annotators, who showed lower agreement in labeling specific hate speech types.

Error analysis reveals key challenges, including false negatives in historically or politically framed speech, and difficulties in detecting subtle dehumanization and coded language. These findings highlight the need for improved models capable of handling implicit bias and contextual nuances. We believe that the dataset and baseline models presented in this study will form the basis for further research on fine-grained hate speech detection.

Limitations

Despite the contributions of HATECAT-TR, several limitations must be acknowledged. First, annotator subjectivity remains a challenge, especially for implicit and context-dependent hate speech. While our semi-automated annotation process improves consistency, some spans remain ambiguous, leading to disagreements among annotators and affecting inter-annotator agreement. This issue is particularly pronounced in the categorization task, where the boundaries between hate speech types can be blurred.

Second, the dataset is limited to Turkish tweets, potentially introducing platform-specific biases. Hate speech manifests differently across social media, news sites, and online forums, requiring broader dataset expansion to improve generalizability. Moreover, some hate speech categories, such as symbolization, are underrepresented, which may affect model performance on more implicit forms of hate speech.

Third, models struggle with detecting subtle, coded, and highly context-dependent hate speech, particularly when expressions carry historical, political, or cultural connotations. Notably, model performance across categories does not always align with annotator performance, suggesting that models rely on different linguistic cues than human annotators. Future work should explore improving contextual understanding, increasing data diversity, and refining annotation guidelines to enhance detection and categorization accuracy.

Finally, Turkish's agglutinative morphology presents specific challenges for span-level detection. Words often contain multiple morphemes, and subword tokenization may not align well with morpheme boundaries. In our comparison, Turkish-specific models produced more coherent segmentations, while multilingual models tended to generate fragmented or unnatural subword units. While tokenization issues alone do not explain the full performance gap, they likely contribute to reduced span accuracy.

Ethical Considerations

This study supports hate speech moderation and research while minimizing potential harm. Since

the dataset contains hateful speech, including offensive or harmful content, reader discretion is advised. To mitigate risks, researchers and annotators were informed in advance about the nature of the content to ensure mental preparedness when handling sensitive material.

To ensure ethical dataset construction, we prioritize privacy and anonymization by removing usernames, preventing individuals from being identified. The dataset is designed to be inclusive, covering hate speech targeting eight minority groups, but some hate speech categories may be underrepresented, and biases in annotation or model performance could disproportionately affect certain groups.

Responsible use of the dataset is essential, as it is intended strictly for academic research. Any misuse, such as attempts to amplify hate speech or evade moderation systems, is strongly discouraged. Given the sensitive nature of hate speech, all annotations underwent manual review to enhance reliability.

While HATECAT-TR and models trained on such data provides a valuable resource for Turkish hate speech detection, responsible and ethical use is crucial. Future work should focus on reducing biases, improving contextual detection, and adapting to the evolving nature of online hate speech to create safer digital spaces.

AI Assistants

AI assistants, ChatGPT[2] and Claude[3], were used in early drafts of this document for proofreading, spell-checking, and grammar correction. The final manuscript has been thoroughly reviewed to ensure correctness and coherence.

Acknowledgments

This work was supported by the EU project "Utilizing Digital Technology for Social Cohesion, Positive Messaging and Peace by Boosting Collaboration, Exchange and Solidarity" (EuropeAid/170389/DD/ACT/Multi), carried out by the Hrant Dink Foundation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

İnanç Arın, Zeynep Işık, Seçilay Kutal, Somaiyeh Dehghan, Arzucan Özgür, and Berrin Yanikoğlu. 2023. Siu2023-nst-hate speech detection contest. In *2023 31st Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Patrick Bareiß, Roman Klinger, and Jeremy Barnes. 2024. English prompts are better for nli-based zero-shot emotion classification than target-language prompts. In *Companion Proceedings of the ACM Web Conference 2024*, pages 1318–1326.

Fatih Beyhan, Buse Çarık, İnanç Arın, Ayşecan Terzioğlu, Berrin Yanikoglu, and Reyyan Yeniterzi. 2022. A Turkish hate speech dataset and detection system. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4177–4185, Marseille, France. European Language Resources Association.

Kaan Büyükdemirci, Izzet Emre Kucukkaya, Eren Ölmez, and Cagri Toraman. 2024. JL-hate: An annotated dataset for joint learning of hate speech and target detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9543–9553, Torino, Italia. ELRA and ICCL.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408, Bangkok, Thailand. Association for Computational Linguistics.

Çağrı Çöltekin. 2020. A corpus of turkish offensive language on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6174–6184.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

Hrant Dink Foundation. 2019. *Hate Speech and Discriminatory Discourse in Media 2019*. HDF Publications.

Hrant Dink Foundation. 2025. *Utilizing AI Against Hate Speech: A Guide to Annotation, Classification, and Detection*. HDF Publications.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

[2] chatgpt.com

[3] claude.ai/chat

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023. ViHOS: Hate speech spans detection for Vietnamese. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 652–669, Dubrovnik, Croatia. Association for Computational Linguistics.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 117–122.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *arXiv preprint arXiv:2312.03732*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Sarah Masud, Manjot Bedi, Mohammad Aflah Khan, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Proactively reducing the hate intensity of online posts via hate speech normalization. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3524–3534.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Itai Mondshine, Tzuf Paz-Argaman, and Reut Tsarfaty. 2025. Beyond english: The impact of prompt translation strategies across languages and tasks in multilingual llms. *Preprint*, arXiv:2502.09331.

Ali Najafi and Onur Varol. 2024. Turkishbertweet: Fast and reliable large language model for social media analysis. *Expert Systems with Applications*, 255:124737.

John Pavlopoulos, Leo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. 2022. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3721–3734, Dublin, Ireland. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Stefan Schweter. 2020. BERTurk - BERT models for Turkish.

Gemma Team et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Cagri Toraman, Furkan Şahinuç, and Eyup Halit Yilmaz. 2022. Large-scale hate speech detection with cross-domain transfer. *arXiv preprint arXiv:2203.01111*.

Gökçe Uludoğan, Somaiyeh Dehghan, Inanc Arin, Elif Erol, Berrin Yanikoglu, and Arzucan Özgür. 2024. Overview of the hate speech detection in Turkish and Arabic tweets (HSD-2Lang) shared task at CASE 2024. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 229–233, St. Julians, Malta. Association for Computational Linguistics.

Linda Zhou, Andrew Caines, Ildiko Pete, and Alice Hutchings. 2023. Automated hate speech detection and span extraction in underground hacking and extremist forums. *Natural Language Engineering*, 29(5):1247–1274.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. HITSZ-HLT at SemEval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 521–526, Online. Association for Computational Linguistics.

A Initial Dataset

The tweets were collected from 2022 to 2024 based on current events and specific hashtags/keywords that focus on groups frequently targeted by hate speech. The dataset spans a diverse range of topics and targets, including tweets related to the Israel–Palestine conflict, anti-Greek sentiment in Turkey, and anti-refugee sentiment. It also includes tweets from the election period, containing hate speech directed at various ethnic and religious identities, such as Arabic, Alevi, Armenian, Kurdish, and LGBTI+ communities.

B Automated Span Annotation

We utilized the following system message to instruct GPT-4o model during annotation:

```
You are a crowd-source worker who
    identifies and labels hate
    speech in Turkish tweets. For
    each tweet, find the hate speech
    spans targeting the given group
    and label them with the hate
    speech category.

Hate Speech Categories:
- Exclusive/Discriminatory Discourse
- Exaggeration; Generalization;
    Attribution; Distortion
- Threat of Enmity; War; Attack;
    Murder; or Harm
- Symbolization
- Swearing; Insult; Defamation;
    Dehumanization
Return a JSON for each tweet:
{
  "tweet": {
    "id": "",
    "span": [
      {
      "text_span": "",
      "hate_speech_category": ""
      },
      {
      "text_span": "",
      "hate_speech_category": ""
      }
    ]
  }
}
```

Listing 1: System message for automated annotation

We used English prompts for Turkish input to leverage LLMs' stronger alignment with English instruction formats. Prior studies have shown that translating prompts into English can improve performance on non-English tasks when using English-centric models (Bareiß et al., 2024; Mondshine et al., 2025; Liu et al., 2025). Given that LLM outputs are sensitive to prompt design, we experimented during pilot testing with extended definitions and examples for each hate speech category. However, these more detailed prompts did not improve annotation quality. As the LLM-generated spans served only as initial suggestions, all outputs were subsequently reviewed and corrected by human annotators, ensuring the reliability of the final dataset. Due to this human verification step, we did not further optimize the prompt beyond the pilot phase.

C Span Annotation Guidelines

C.1 Annotators

The annotators were recruited as part of an ongoing project and had previously received training from the Hrant Dink Foundation (HDF) on hate speech detection, including its identification and categorization. The team consisted of six annotators with academic backgrounds in various disciplines: two from linguistics, three from cultural studies, and one from political science. They were either students involved in projects or employees of the organization. Compensation details were aligned with the HDF's standard policies for research assistants and project contributors.

C.2 Annotation Process

To ensure high-quality span annotations, additional refinement was performed through manual review. All GPT-4o-generated spans were manually reviewed by human annotators, with each tweet assigned to two different annotators. Our process involved evaluating all spans for each tweet rather than relying on random subsets per category. The annotators were randomly assigned tweets, ensuring that each tweet was independently reviewed by two individuals. Annotations exhibiting substantial disagreements regarding hatefulness or with an Intersection over Union (IoU) score below 0.5 between annotators were subjected to a secondary review. A third annotator subsequently conducted a secondary review, selecting the most appropriate annotation from the two initial reviewers or rejecting both if neither was satisfactory.

The semi-automated approach allowed annotators to verify and revise existing annotations rather than annotate from scratch. Without this method, they would have manually labeled all 2981 tweets, including the 1530 that were ultimately accepted without changes, which would have significantly increased the workload. Instead, annotators focused only on revising 1451 tweets, addressing span boundaries, correcting category labels, or removing hallucinated content. Given the complexity of identifying spans and assigning one of five categories, this approach greatly reduced the annotation effort.

C.3 Annotation Task Definition

The objective of this task is to identify and label the hateful text span (i.e., the text fragment containing hate speech) in tweets targeting specific groups, in-

cluding Greek, Armenian, Kurdish, Alevi, LGBT, immigrant/refugees, Arab, and Jewish communities. For each span identified, annotators must assign the appropriate hate speech category based on its content.

C.4 Pre-Annotated Predictions

Annotators are provided with initial span predictions and category labels generated by GPT-4o to facilitate the annotation process. These pre-annotations should be reviewed and adjusted as necessary: annotators may delete, create, or modify spans to accurately capture hate speech and assign appropriate categories.

C.5 Hate Speech Categories and Examples

Annotators are provided with hate speech categories along with their examples.

Exclusive/Discriminatory Discourse

- **Definition:** Language that excludes or differentiates negatively based on group membership.
- **Tweet Example:** "Armenians do not belong in these lands, they are not one of us."
- **Hateful Span:** "Armenians do not belong in these lands"
- **Category:** Exclusive/Discriminatory Discourse

Exaggeration, Generalization, Attribution, Distortion

- **Definition:** Making unfair generalizations or attributing negative traits to a group.
- **Tweet Example:** "All Kurds disturb the peace, that's why we don't want them in our cities."
- **Hateful Span:** "All Kurds disturb the peace"
- **Category:** Exaggeration; Generalization; Attribution; Distortion

Threat of Enmity, War, Attack, Murder, or Harm

- **Definition:** Statements suggesting or inciting violence or harm toward a group.
- **Tweet Example:** "We should get rid of these Jews, we can't tolerate it anymore."
- **Hateful Span:** "We should get rid of these Jews"
- **Category:** Threat of Enmity; War; Attack; Murder; or Harm

Symbolization

- **Definition:** Using symbols or traits to reinforce negative stereotypes about a group.
- **Tweet Example:** "How can a Kurdish deputy represent us? They can't work for this country."
- **Hateful Span:** "How can a Kurdish deputy represent us?"
- **Category:** Symbolization

Swearing, Insult, Defamation, Dehumanization

- **Definition:** Direct insults or dehumanizing portrayals of a group.
- **Tweet Example:** "These Syrian scum have ruined our country."
- **Hateful Span:** "These Syrian scum"
- **Category:** Swearing; Insult; Defamation; Dehumanization

C.6 Annotation Interface and Guidelines

The annotation process is conducted using Label-Studio[4]. Figure 2 illustrates the annotation interface used for labeling hate speech spans in tweets. Key components of the annotation interface include:

- **Green Box:** Displays annotation versions (v1 for the model's prediction and for annotators' versions).
- **Black Box:** Highlights the target group affected by the hate speech.
- **Blue Box:** Contains category buttons for annotators to assign labels.
- **Red Box:** Displays the text to be annotated.
- **Yellow Box:** Shows all annotated spans, while the **orange box** highlights the currently selected annotation.

Annotators should first select a category button before highlighting the hateful span. The red trash button allows deletion of incorrect annotations.

D Hate Speech Span Mismatch Review Guidelines

D.1 Review Task Definition

Each tweet in this task contains two different hate speech span annotations. The objective is to determine the correct annotation. Annotators should:

- Select the correct annotation if one of the two is accurate.
- Reject both annotations if neither is correct.
- Refer to the hate speech categories and examples for guidance.

[4]https://github.com/HumanSignal/label-studio/

Category	Example Text	English Translation
Exclusive/Discriminatory Discourse	"Türk'ün devletinde biri Gürcü diğeri sözüm ona Kürt iki aday arasında tercih yapmak zorunda kaldık"	"In the Turkish state, we had the choice between a Georgian and the other so-called Kurdish candidates"
Exaggeration; Generalization; Attribution; Distortion	"Kanı ermeni olanlara anlatamazsınız! Çekinmenize göre yok."	"You can't explain it to those with Armenian blood! There is no need to hesitate."
Threat of Enmity; War; Attack; Murder; or Harm	"7 düvel değil, 70 düvel de gelseniz Türklerin yumruğunu hissettiğinizde dünyanızı sapıtmış olursunuz. Anadolu hiçbir zaman Yunan olmadı, Yunan diye bir medeniyet hiç olmadı.	"Not just seven nations, but even seventy nations—when you feel the fist of the Turks, your world will be turned upside down. Anatolia was never Greek, and there was never a Greek civilization. "
Symbolization	"yunan ve ermeni tohumlarını siler atarız"	"We will wipe out the greek and armenian offspring."
Swearing; Insult; Defamation; Dehumanization	"15 mayısta hesaplaşacağız diyen ermeni köpekler vardı"	"On May 15th, there were armenian dogs saying 'we will settle accounts'"

Table 3: Categories of harmful content with examples. Highlighted spans show span annotations.

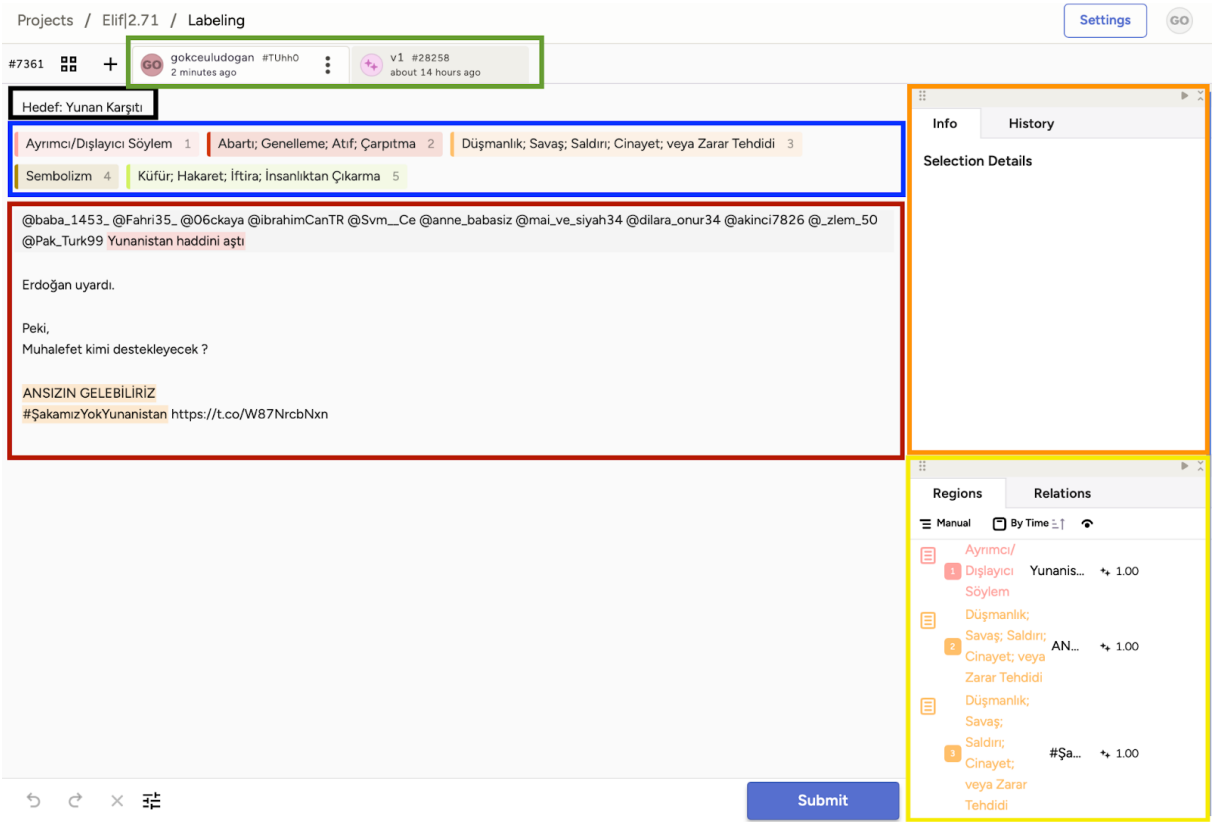

Figure 2: Annotation interface used in the labeling process. The interface includes predefined hate speech categories, model-generated initial annotations, and manual annotations for review. Tweets in Turkish are shown in this example, where spans such as "Yunanistan haddini aştı" ("Greece has crossed the line") are labeled under relevant categories.

D.2 Reviewing Interface

The review process is also conducted in LabelStudio. Figure 3 show the annotation interface used for reviwing annotations, with the following interface elements:

- **Yellow Box:** Displays the targeted group.
- **Red Box:** Lists possible hate speech categories.
- **Green Box:** Contains all annotated spans.
- **Blue Box:** Allows selection between the two annotations or rejection of both.

E Target group distribution

The distribution of the target groups is presented in Table 4.

Table 4: Statistics of the annotated dataset with respect to target groups.

Target Group	Number of Tweets Before Mismatch Handling	Number of Tweets After Mismatch Handling
Jew	1132	925
Greek	1119	1015
Armenian	628	534
Arab	337	301
Alevi	127	108
Kurdish	63	52
LGBTI+	49	46

F Inter-annotator agreement

The overall and category-specific inter-annotator agreements are presented in Table 5.

Metric	F1
Overall Detection	0.77
Overall Categorization	0.50
Exclusive/Discriminatory Discourse	0.44
Threat of Enmity; War; Attack; Murder; or Harm	0.66
Symbolization	0.29
Exaggeration; Generalization; Attribution; Distortion	0.44
Swearing; Insult; Defamation; Dehumanization	0.54

Table 5: Inter-annotator Agreements

G Implementation Details

We formed our test set by randomly selecting 300 tweets from tweets that reached agreement after the first human review stage to ensure its reliability. The remaining tweets were used for training with 5-fold cross-validation. We utilized BERTurk[5], TurkishBERTweet [6], LLaMa-3.3-70B-Instruct[7], LLaMa-3.1-8B[8], and Gemma-2-9B[9] models.

For BERTurk and TurkishBERTweet models, we used the AdamW optimizer with the following hyperparameters: the learning rate is 0.00005, the batch size is 4, the number of epochs is 5, the early stopping patience is 3, and the warm-up ratio is 0.1.

For LLaMa-3.1-8B and Gemma-2-9B, we applied a parameter efficient finetuning with rsLoRA (Kalajdzievski, 2023) and used the AdamW 8 bit optimizer with the following hyperparameters: the learning rate is 0.0002, the batch size is 8, the number of epochs is 1. During the categorization task, Gemma consistently produced unparsable outputs. To address this, we experimented with various settings, including different number of epochs, reduced batch sizes of 2 and 4. as well as 4-bit and 8-bit quantization. Despite these efforts, the issue remained unresolved, and thus we were unable to report Gemma's performance on the categorization task.

We conducted zero-shot and few-shot experiments with with LLaMa-3.3-70B-Instruct. For zero-shot inference, we used the prompt described in Appendix B. For the few-shot setting, the same prompt was extended with additional examples taken from Appendix C.5.

We performed training using 5-fold cross-validation and reported the mean and standard deviation of all runs. Separate models were trained for hate speech detection and categorization tasks.

Evaluation was performed using the SeqEval[10] library. Predictions and true labels are flattened, and precision, recall, and F1 scores are computed.

H Hate Speech Categorization Performance

Table 6 presents F1 scores for five hate speech categories across a range of models.

[5]https://huggingface.co/dbmdz/bert-base-turkish-cased

[6]https://huggingface.co/VRLLab/TurkishBERTweet

[7]https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[8]https://huggingface.co/meta-llama/Llama-3.1-8B

[9]https://huggingface.co/google/gemma-2-9b

[10]https://github.com/chakki-works/seqeval

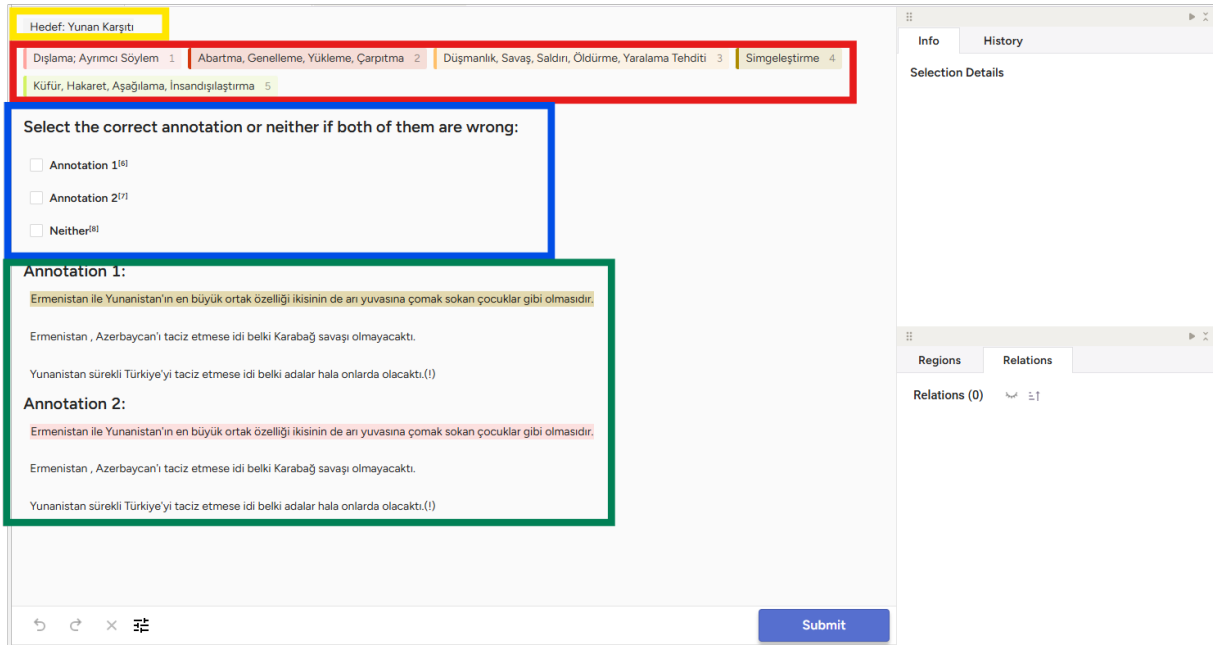

Figure 3: Review interface used in the annotation process. The example displays a tweet in Turkish where reviewers must select the correct annotation or reject both if necessary. Highlighted spans correspond to the identified hate speech segments.

Category	**BERTurk**	**TurkishBERTweet**	**LLaMa-3.3-70B Few Shot**	**LLaMa-3.3-70B Zero Shot**	**LLaMa-3.1-8B Finetuning**
Exclusive/Discriminatory Discourse	0.08 ± 0.03	0.02 ± 0.03	0.12 ± 0.01	0.09 ± 0.01	0.07 ± 0.04
Exaggeration; Generalization; Attribution; Distortion	0.17 ± 0.07	0.10 ± 0.04	0.08 ± 0.02	0.08 ± 0.01	0.02 ± 0.02
Threat of Enmity; War; Attack; Murder; or Harm	0.44 ± 0.02	0.35 ± 0.02	0.33 ± 0.01	0.23 ± 0.01	0.10 ± 0.02
Symbolization	0.22 ± 0.05	0.21 ± 0.06	0.00 ± 0.00	0.00 ± 0.00	0.09 ± 0.04
Swearing; Insult; Defamation; Dehumanization	0.31 ± 0.02	0.29 ± 0.04	0.26 ± 0.01	0.28 ± 0.01	0.13 ± 0.04

Table 6: Hate speech categorization performance.